The Multilevel Approach to Repeated Measures for Complete and Incomplete Data

Maas, CJM; Snijders, TAB

Published in:
Quality & Quantity

Publication date:
2003

Quality & Quantity **37:** 71–89, 2003.

The Multilevel Approach to Repeated Measures for Complete and Incomplete Data

CORA J. M. MAAS[1] and TOM A. B. SNIJDERS[2]
[1]*Department of Methodology and Statistics, University of Utrecht;* [2]*Department of Statistics and Measurement Theory, University of Groningen*

Abstract. Repeated measurements often are analyzed by multivariate analysis of variance (*MANOVA*). An alternative approach is provided by multilevel analysis, also called the hierarchical linear model (*HLM*), which makes use of random coefficient models. This paper is a tutorial which indicates that the *HLM* can be specified in many different ways, corresponding to different sets of assumptions about the covariance matrix of the repeated measurements. The possible assumptions range from the very restrictive compound symmetry model to the unrestricted multivariate model. Thus, the *HLM* can be used to steer a useful middle road between the two traditional methods for analyzing repeated measurements. Another important advantage of the multilevel approach to analyzing repeated measures is the fact that it can be easily used also if the data are incomplete. Thus it provides a way to achieve a fully multivariate analysis of repeated measures with incomplete data.

Key words: *MANOVA*, incomplete data, missing at random, hierarchical linear model, Hotelling's test, Wald test, compound symmetry model.

1. Introduction

Repeated measures data are common in many disciplines. Procedures for analysing such data are treated, e.g., in O'Brien and Kaiser (1985), Maxwell and Delaney (1990), and Stevens (1996). In the period before 1985, mainly the compound symmetry model and the closely related sphericity model were used. The compound symmetry model represents the dependence between the several data obtained from a single individual by a random main effect of the individual. The paper by O'Brien and Kaiser (1985) marks the transition to the use of procedures based on multivariate analysis of variance (*MANOVA*). In the *MANOVA* model, no assumptions are made about the covariance matrix of the repeated measurements. The only assumptions are independence and identical distributions within treatment groups, homoscedasticity between groups, multivariate normality, and complete data. The last assumption means that, if there are p measurement occasions, for each subject in the data set the measurements on all p occasions are available.

Since the seminal paper by Laird and Ware (1982), random coefficient models, or linear mixed models, have been increasingly used for analysing repeated measurements. These models have also been used in multilevel analysis (Bryk and Raudenbush, 1992; Goldstein, 1995; Snijders and Bosker, 1999), a methodology

for the analysis of clustered data in general. Repeated measurements are one type of clustered data (measurements clustered within individuals), but other types of clustering (e.g., pupils within classes, classes within schools; clients within therapists; voters within voting precincts) are also frequent. In the multilevel literature, the hierarchical linear model (*HLM*) is the term used for the linear mixed model with nested random coefficients. The multilevel analysis of repeated measures is treated in, e.g., Bryk and Raudenbush (1992, Chap. 6), Goldstein (1995, Chaps. 4 and 6), Rogosa et al. (1982), Snijders and Bosker (1999, Chap. 12), and Van Der Leeden (1998).

The present paper explains the relations between the random coefficient, or multilevel, approach to repeated measurements, and the traditional treatments based on the compound symmetry model and the *MANOVA* model, paying special attention to incomplete data. The specification of the random coefficient model will be shown to imply specific assumptions for the covariance matrix of the p repeated measurements, with the compound symmetry model and the unrestricted *MANOVA* model as special cases. Understanding the assumptions implied by the random coefficient model is important, because the analysis may lead to erroneous conclusions if these assumptions are not satisfied.

The reformulation of the unrestricted *MANOVA* model as a multilevel model is not usual (although by no means new; cf. Goldstein, 1995), and therefore special attention is given to this formulation. An important advantage of the multilevel approach is that incompleteness of the data on the dependent variable does not complicate the analysis, provided that missingness is at random. The missing observations simply can be omitted from the data set. This implies that the multilevel approach allows the analysis of incomplete repeated measures data without restrictive assumptions on the covariance matrix. However, the multivariate F-tests of the *MANOVA* approach, which are exact for the case of one or two groups, are replaced in the multilevel approach by likelihood ratio (deviance) or by Wald tests, which rely on large sample approximations. It is explained below for some basic repeated measures designs how the Wald tests correspond to the multivariate F-tests.

2. Missing Observations in Repeated Measures

Incompleteness of data is common in empirical research. In this paper we consider only missingness of the dependent variable, because in repeated measures analysis the independent variables usually are completely observed. An essential question is about the mechanism that causes incompleteness of data, called here the *response process*. A classification of the missing-data mechanism can be made according to how the response process depends on the observed and unobserved values (Little and Rubin, 1987). If the response process is independent of all variables, whether observed or missing, the missing data are "missing completely at random" (*MCAR*). Loosely speaking, if the response process depends on observed but not on unobserved variables, the data are "missing at random" (*MAR*). A more formal

way of expressing this is that the conditional probability of observing the response, given all variables (observed as well as unobserved), is the same as the conditional probability of observing the response, given only the observed variables. A generic example of MCAR is incompleteness due to randomly failing apparatus. Examples of MAR are recording failures depending on group or measurement occasion, but not otherwise on the (unrecorded) value; and termination of the observations after recording a value above or below a given threshold. An example of non-MAR is drop-out from a therapy comparison study as a consequence of recovery or of psychological breakdown without this being predictable from the data that were observed earlier. In the cases of MAR or MCAR data, valid likelihood-based statistical inference is possible without modeling the response process (Little and Rubin, 1987; Schafer, 1997). If the probability of response depends on the unobserved dependent variable (and not only as a function of the independent variables), the missingness itself is informative, and it is preferable to specify a model that employs this information.

2.1. TRADITIONAL TREATMENTS OF MISSING DATA

The multivariate F-tests of the *MANOVA* approach to repeated measures require a complete data matrix. When data are incomplete, and MAR is a reasonable assumption, researchers often choose either of the following options:

1. all cases with any missing values are removed ("listwise deletion");
2. the missing data are estimated ("imputed", cf. Little and Rubin, 1987) and analysis for a complete data set is performed.

In the first situation, valuable information is lost. In the second situation, the question is: "how to estimate the missing data" and "how good are these estimates and the resulting tests". Better than single imputation methods is multiple imputation, cf. Rubin (1987) and Schafer (1997).

Procedures for estimating repeated measures with incomplete data and an unrestricted covariance matrix are discussed, a.o., by Berk (1987), Jennrich and Schluchter (1986), Little and Rubin (1987), and Schafer (1997). Procedures for estimating repeated measures with incomplete data and a structured covariance matrix were proposed by Laird and Ware (1982) and Jennrich and Schluchter (1986). Such procedures are now incorporated in *SAS Proc Mixed* (Littell, Milliken et al., 1996) and in BMDP-5V (Dixon, 1992).

The multilevel packages *MlwiN* (Goldstein et al., 1998) and *HLM* (Raudenbush et al., 2000) can also be used to obtain estimates and tests for repeated measures data under various specifications for the covariance matrix for complete as well as incomplete data. (For the *MlwiN* commands see Snijders and Maas, 1996). The tests available for incomplete data are likelihood ratio ('deviance') and Wald tests, which are large-sample tests respecting the level of significance only approximately, whereas the F-tests and multivariate F-tests for complete data are exact for homoscedastic multivariate normal data.

3. Basic Assumptions

Throughout this paper we assume that p treatments (the levels of one within-subjects factor) are being tested in a repeated measures design. In other words, there are p measurement occasions and N subjects, and subject i provides observation Y_{ij} on measurement occasion, or treatment, or condition, j. In the case of complete data, each subject provides p measurements, Y_{i1} to Y_{ip}, combined in the vector Y_i . The standard assumption is that Y_i has a multivariate normal distribution. Under this assumption, the relevant parameters of the distribution of Y_i are the vector of means and the covariance matrix. The tested hypotheses usually refer to the vector of means, the covariance matrix playing the role of a nuisance parameter. The population mean of Y_{ij} is denoted by μ_j.

In the compound symmetry model, which is the basis of the classical "univariate" approach to repeated measures (see, e.g., Maxwell and Delaney, 1990, Chap. 11), the data are represented as the sum of the population mean for occasion j, the subject main effect U_i, and random error:

$$Y_{ij} = \mu_j + U_i + E_{ij}. \tag{1}$$

The compound symmetry model implies that all variances have a common value and all covariances have a common value. By contrast, in the multivariate approach (e.g., Maxwell and Delaney, 1990; Stevens, 1996) no assumptions are made with respect to the covariance matrix of Y_i .

4. The Multilevel Formulation of Repeated Measures Data

The multilevel model (or Hierarchical Linear Model) for two levels (where level 1 is understood to be nested in level 2) is composed of three parts: the *fixed part*, representing fixed effects; and the *random part at level 1* and the *random part of level 2*, representing unexplained variability. The multilevel model is a special case of the mixed model (e.g., Hays, 1988), and is also called a random coefficient model. In the multilevel representation of repeated measures, the measurement occasions constitute the first level and the individuals the second. The data are represented in a format where each "case", or record, is identified by its subject ("level-2 unit") i and its measurement occasion ("level-1 unit") j, so that to each case there belongs a single measurement Y_{ij}. (In the literature on multilevel analysis the converse notation is usual, with j indicating the level 2 unit and i indicating the level 1 unit. In this paper we stick to the conventional repeated measures notation.) Data are represented in this way also in the classical "univariate" approach to repeated measures (e.g., Maxwell and Delaney, 1990, Chap. 11). The fixed part defines the vector of means and the random part defines the covariance matrix.

The random coefficient approach to repeated measures is often, but not necessarily, based on polynomial trend models. If the time moment associated with measurement occasion j is denoted by t_j , such models are based on functions

$f_h(t_j)$ $(h = 1, \ldots, H)$ defined as polynomial functions of time. Function f_h is then a polynomial of degree $h - 1$ and $f_1 = 1$ is constant. The functions can be defined by $f_{hj} = f_h(t_j) = t_j^{h-1}$ but also by a set of orthogonal polynomials. Depending on the research topic one may use any other linearly independent set of functions f_h instead of polynomials. A random coefficient trend model then is defined as

$$Y_{ij} = \sum_{h=1}^{H} \pi_h f_{hj} + \sum_{h=1}^{q} U_{hi} f_{hj} + E_{ij}, \tag{2}$$

with, usually, $q \leq H$. The parameters π_h are parameters defining the vector of means. In a between-subjects design, π_h also depends on the group (between-subjects factor). Associated to individual i are the correlated random variables U_{1i} to U_{qi}. These random variables are defined at level 2, the level of the individual subject. The assumption for the N vectors $U_i = (U_{i1}, \ldots, U_{iq})$ is that they are independent and identically distributed across subjects, having a multivariate normal distribution with mean $\mathbf{0}$ and arbitrary covariance matrix. Finally, E_{ij} is a random error with population mean 0 and variance σ_E^2.

This model can be regarded as a decomposition of the measurements Y_{ij} into an individual-specific curve,

$$\sum_{h=1}^{H} \pi_h f_{hj} + \sum_{h=1}^{q} U_{hj} f_{hj},$$

plus a random error E_{ij} (the random part at level 1). The individual-specific curve is the sum of a mean population curve

$$\sum_{h=1}^{H} \pi_h f_{hj}$$

(the fixed part of the model) and an individual deviation

$$\sum_{h=1}^{q} U_{hi} f_{hj}.$$

(the random part at level 2 which represents inter-individual variation). The individual-dependent coefficient U_{hi} is the *random component of the slope* of the level-one variable f_h. The variables f_1 to f_q have fixed as well as random effects, whereas f_{q+1} to f_H have only fixed effects.

A saturated (i.e., unrestricted, and perfectly fitting) model for the fixed part is obtained when the number of terms H is equal to the number of observations p

and the coefficients π_h are unrestricted. In that case there is a one-to-one correspondence between the vector of occasion means $(\mu_1, \ldots, \mu_p)$ and the vector of coefficients $(\pi_1, \ldots, \pi_p)$, defined by the equations

$$\mu_j = \sum_{h=1}^{p} \pi_h \quad (j = 1, \ldots, p).$$

It is a matter of convenience whether one prefers the parametrisation by μ or by π.

4.1. COVARIANCE MATRICES IMPLIED BY THE MULTILEVEL MODEL

Let us assume for a moment that a saturated model is indeed used for the fixed part. Then the specification of the random coefficient model amounts to the choice of the number q and the functions $f_1, \ldots, f_q$; this choice implies the selection of a set of possible covariance matrices. The simplest choice is $q = 1$ with the constant value $F_{1ij} = 1$ (all i, j). This yields exactly the compound symmetry model defined in (1). In multilevel terminology this is the random intercept model. Another possibility is to have $q = 2$ random coefficients and define $f_{1ij} = 1$, $f_{2ij} = t_j$. This is a trend model where the individuals follow individual curves which deviate from the population average by a constant term and a linear trend, both being individual-specific. If the fixed part is represented by occasion means μ_j, this random slope model is given by

$$Y_{ij} = \mu_j + U_{1i} + U_{2i} t_j + E_{ij}, \tag{3}$$

where the random intercept U_{1i} and the random slope U_{2i} are allowed to be correlated. More generally, any number $q < p$ of polynomial (or other) functions of time can be given correlated random slopes.

Denote the variances and covariances of the repeated measurements by $\sigma_{jj} = \sigma_j^2 = \text{var}\,(Y_{ij})$ and $\sigma_{jk} = \text{cov}\,(Y_{ij}, Y_{ik})$. Then model (3) with the single random slope implies the covariance matrix defined by

$$\sigma_{jk} = \tau_1^2 + (t_j + t_k)\tau_{12} + t_j t_k \tau_2^2 + \delta_{jk}\sigma_E^2,$$

(for $j, k = 1, 2$) where τ_1^2, τ_2^2, and τ_{12} denote the variances and covariance of the random coefficients U_{1i} and U_{2i} and $\delta_{jk} = 1$ for $j = k$, and 0 otherwise. Generally, for q random slopes,

$$\sigma_{jk} = \sum_{h,m=1}^{q} f_{hj} f_{mk} \tau_{hm} + \delta_{jk}\sigma_E^2.$$

Thus, random coefficient model (2) describes the covariance matrix of the p repeated measurements by means of $q(q+1)/2 + 1$ parameters, viz., the covariance matrix of $(U_{1i}, \ldots, U_{qi})$ plus the variance of E_{ij}.

Thus we see that random coefficient models with one or more random coefficients imply restrictions on the covariance matrix which are intermediate between the compound symmetry model and the fully multivariate model. E.g., the compound symmetry model has two parameters for the covariance matrix (viz., τ_1^2 and σ_E^2), model (3) with one random slope has four parameters, but the total covariance matrix has $p(p+1)/2$ independent entries which all are free parameters in the fully multivariate model.

The lowest number of individual-dependent random terms is $q = 0$, corresponding to independent measurements; the next is $q = 1$, for the compound symmetry model; the highest is $q = p - 1$, because $q = p$ would lead to $p(p+1)/2 + 1$ parameters, one more than the total number of free parameters in the covariance matrix. For $q = p - 1$, however, the model for the covariance matrix is not yet saturated, since it still has $p(p+1)/2 - \{p(p-1)/2 + 1\} = p - 1$ parameters less than the full covariance matrix. An unrestricted model can be obtained by using p terms in the random part at level 2 and omitting the random part at level 1. The resulting model is

$$Y_{ij} = \sum_{h=1}^{H} \pi_h f_{hj} + \sum_{h=1}^{p} U_{hi} f_{hj} \tag{4}$$

with covariance matrix

$$\sigma_{jk} = \sum_{h,m=1}^{p} f_{hj} f_{mk} \tau_{hm}. \tag{5}$$

It was mentioned above that, in principle, any set of linearly independent functions of j can be used for $f_1, \ldots, f_q$. For the unrestricted model (4), where $H = q = p$, one convenient possibility proposed by Goldstein (1995, Chap. 4) is to use dummy coding defined by

$$z_{hj} = \begin{cases} 1 & h = j \\ 0 & h \neq j. \end{cases} \tag{6}$$

In words, p dummy variables are used, one for each measurement occasion. These variables have correlated random slopes at level 2, and the constant term (normally used for the random intercept) is not used. With this dummy coding, the covariance matrix of the observed variables is equal to the covariance matrix of the random terms U_{hi}, i.e., $\sigma_{jk} = \tau_{jk}$ in (5).

It can be concluded that the multilevel, or random coefficient, model can represent a variety of covariance matrices, ranging from the complete independence and compound symmetry models to the fully multivariate model.

4.2. FITTING A COVARIANCE MATRIX

If not enough parameters are used to model the covariance matrix, the risk is that this matrix is wrongly specified. Tests about the vector of means may be incorrect. On the other hand, these tests will have unnecessarily low power if too many parameters are used to represent the covariance matrix (cf. Reinsel, 1984 and Ware, 1985). It is important to have a well-fitting model for the covariance matrix, but – especially for small and intermediate sample sizes – to refrain from overfitting. The literature accordingly contains warnings against using the compound symmetry model unless one is confident that its assumptions are satisfied (Maxwell and Delaney, 1990; O'Brien and Kaiser, 1985).

In the context of the random coefficient model (2) for the covariance matrix, this means that, given a meaningful choice of the functions f_h , one should choose q large enough but not too large. A good procedure is first to choose an adequate model for the covariance matrix while using a saturated model (i.e., $H = p$) for the vector of means; with the resulting fitted model for the covariance matrix, one can proceed to test the interesting hypotheses about the vector of means. To determine an adequate model for the covariance matrix, the fully multivariate model (4) and random coefficient models (2) with different values of q can be compared by means of stepdown likelihood ratio (also called deviance) tests described below. Using the deviance (defined as minus twice the maximized log-likelihood) of the unrestricted model (4) as the point of departure, one can go stepwise to model (2) and determine the lowest number q of random coefficients for which (2) still yields an acceptable fit. The stepdown tests are carried out by comparing the model having q random slopes with the model having $q-1$ random slopes, starting from $q = p$, decreasing q by 1 when the result is non-significant and stopping at the first significant result. The smallest value of q which yielded a non-significant test is chosen for the fitted model. If to express the true covariance matrix in the population a minimum of q_0 random slopes is required, then the probability that this stepdown procedure leads to a value smaller than q_0 (i.e., an error of the first kind is made in the total procedure) is less than or equal to the significance level used in the separate stepdown tests.

4.3. DIFFERENTLY STRUCTURED COVARIANCE MATRICES

There are other families of covariance matrices which also provide a middle road between the compound symmetry model and the fully multivariate. Within the multilevel framework, the model can be extended by letting the variance of E_{ij} depend on j (see Chapter 8 of Snijders and Bosker, 1999, on heteroscedasticity). Other models are outside the multilevel framework. A well-known model is the autoregressive model, where each measurement is regressed on preceding measurements. A stationarity first order autoregressive model has a correlation matrix of the form $\sigma_{jk} = \sigma_0^2 \rho^{|j-k|}$ where ρ is the autocorrelation coefficient. Another model is the moving average model, where the covariance matrix is banded, i.e.,

σ_{jk} depends arbitrily on the "time lag" $|j-k|$ for $|j-k|$ smaller than some number q, the order of the moving average, and is 0 for larger values of $|j-k|$. These two models are often used in time series analysis.

4.4. INCOMPLETE DATA

A nice thing about these "univariate-type" formulations is that they are not affected by missing data. Complete data are represented by a data set with np records, each record containing the three numbers i, j, and Y_{ij}, and codes for any between-subject factors. Incomplete data are simply represented by a data set with less records. Since the data is represented in (2) and (4) not by a vector but by sums involving fixed and random coefficients, the representation is the same whether data are complete or incomplete. The number of available measurements per subject may range from 1 to p. Estimation methods and tests can be applied to incomplete as well as to complete data, with the caveat that the standard errors and significance levels rely on asymptotic approximations, in contrast to the exact tests available for complete data for the compound symmetry model and the fully multivariate model with one or two groups.

The F-tests for the compound symmetry model also are valid under the weaker assumption of sphericity, defined by the requirement that a complete set of orthonormal contrasts has constant variances and zero correlations (see, e.g., Maxwell and Delaney, 1990, or Stevens, 1996). This was generalized to incomplete data by Schwertman (1978).

5. Estimation and Testing

There are two major estimation procedures for random coefficient models: maximum likelihood (ML) and residual (or restricted) maximum likelihood (REML) estimation (see Bryk and Raudenbush, 1992 or Goldstein, 1995). REML estimation takes into account, in the estimation of the parameters of the random part, the loss of degrees of freedom resulting from the estimation of the parameters of the fixed part. This has an indirect effect on the estimates of the fixed part. For large sample sizes these two estimation procedures do not differ much.

In the case of incomplete data, the usual "naive" procedure is to estimate means, variances, and covariances on the basis of the available data (with pairwise deletion for covariances). The ML and REML estimates of the parameters differ from these estimates (mostly only slightly, unless there are many missing data and correlations are high) and are statistically more efficient, as may be deduced from Little and Rubin (1987).

Hypothesis tests can be based on the *deviance*, defined as minus twice the natural logarithm of the maximized likelihood (which is the probability density function, filling in the estimated parameter values). The deviances can be used for hypothesis tests. If two models M_0 and M_1 are compared, where M_0 has the role of

the null hypothesis and is a sub-model of M_1, the deviance difference between the two models is the test statistic, which if M_0 is true has an asymptotic chi-squared distribution, the number of degrees of freedom being the additional number of free parameters in M_1 as compared to M_0 .

Another test is the Wald test (described for general models, e.g., in Rao (1973, section 6e); for multilevel models see Bryk and Raudenbush (1992, formula [3.73]) or Snijders and Bosker (1999, formula (6.4))), which can be applied to test an arbitrary vector of linear combinations of the fixed parameters. The Wald statistic for an r-dimensional subparameter θ is defined by

$$W = \hat{\theta}' \hat{\Sigma}_{\hat{\theta}}^{-1} \hat{\theta} \tag{7}$$

where $\hat{\theta}$ is the ML estimate for θ and $\hat{\Sigma}_{\hat{\theta}}$ is the ML estimate for the covariance matrix of the parameter estimate $\hat{\theta}$. The Wald statistic has asymptotically (for large sample sizes), if $\theta = 0$, a chi-squared distribution with r degrees of freedom. It is also possible to base the Wald test on the REML estimates. This yields slightly different but asymptotically equivalent results.

Example 1

Hand and Taylor (1987) present data from a study of the effects of drinking alcohol on salsolinol excretion. Subjects are $N = 14$ alcohol dependent individuals, divided into two groups (group 1 is moderately, group 2 severely alcohol dependent). There are $p = 4$ consecutive days of measurement. The dependent variable is logarithmically transformed salsolinol concentration in a urine sample. The 14 subjects provided complete data. These data are used in the examples throughout this paper. The F-tests can be calculated for complete data by software implementing the *MANOVA* repeated measures model like *SPSS*, *SAS*, and *BMDP*. The Wald tests and deviance tests can be calculated by software for random coefficient models like *HLM*, *MlwiN*, and *SAS*. Random coefficient models were fitted with a saturated fixed part and random parts defined by polynomial trends as in (2), where $f_{hj} = (j - 2.5)^{(h-1)}$. Table 1 contains the deviances and the number of parameters for the covariance matrix for the fully multivariate model ($q = 4$) and for the model with $q = 0, 1, 2, 3$ trend terms in the random part at level 2, not using the division into two groups.

The deviances for $q = 2$ and 3 are equal, and the same holds for $q = 0$ and 1. This happens occasionally, and means that for $q = 1$ and 3, the maximum likelihood estimate occurs at a boundary point of the parameter space where one of the variance parameters equals 0. It is no reason for concern.

Since the sample size is small and the purpose of this part of the analysis is to arrive at a well-fitting covariance matrix, it is reasonable to use a significance level of 0.10 for the stepdown tests. The first stepdown test has $\chi^2 = 135.46 - 127.81 = 7.65$, $d.f. = 10 - 7 = 3$, $p = 0.054$. At a significance level of 0.10, the fully multivariate model fits better than the model with $q = 3$ random slopes, and therefore is retained for the analysis of the fixed part.

Table I. Deviance values for the covariance matrix of the salsolinol data with unrestricted fixed part. p-values are for the chi-squared test comparing the models for q and $q+1$.

q	# parameters	deviance	p
4	10	127.81	–
3	7	135.46	0.054
2	4	135.46	1.00
1	2	139.36	0.14
0	1	139.36	1.00

If the stepdown tests would have used the significance level of 0.05, the first stepdown test would be passed, just like all subsequent stepdown tests. Then the final model would be the model with $q = 0$, i.e., the complete independence model.

5.1. TESTING HOMOGENEITY OF MEANS

Consider the one-group design, which has no between-subject factors. A basic null hypothesis is the equality across treatments of the mean responses, expressed by

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p.$$

The usual procedure in the *MANOVA* approach is to transform the p dependent variables to $p-1$ contrasts, e.g., difference contrasts $D_{ij} = Y_{i,j+1} - Y_{ij}$, combined into the vector $D_i = (D_{il}, \ldots, D_{i,p-1})$. The null hypothesis of homogeneity of the means of Y_{ij} corresponds to the hypothesis that the population mean of D_i is $\mathbf{0}$. This hypothesis is tested by means of Hotelling's T^2 test statistic (Anderson, 1984; Stevens, 1996). The test statistic is defined by

$$T^2 = N\bar{\mathbf{D}}'\mathbf{S}_D^{-1}\bar{\mathbf{D}},$$

where $\bar{D}$ is the mean and $\mathbf{S}_D$ the observed covariance matrix of D_i $(i = 1, \ldots, N)$. To test the significance, an exact F transformation of T^2 is given by

$$F = \frac{N - p + 1}{(N-1)(p-1)} T^2,$$

with $(p-1)$ and $(N-p+1)$ degrees of freedom. The choice of the $p-1$ contrasts does not affect the value of T^2, provided that the contrasts are linearly independent.

The null hypothesis of homogeneity of means is represented in (2) or (4) by $H = 1$. The only variable with a fixed effect then is the constant (always equal to

1). Thus we see that the analysis of repeated measures leads to multilevel models where some variables have a random but not a fixed effect, which is unusual in most applications of multilevel modeling.

In the multilevel approach to repeated measures, the transformation to the contrasts D_{ij} is superfluous, because the null hypothesis is expressed by (2) or (4) as a model for the original variables Y_{ij} rather than for the contrasts.

Suppose that homogeneity of means is tested within the context of the fully multivariate model (4). If the data are complete, the deviance (likelihood ratio) test statistic is a non-linear increasing function of Hotelling's T^2 (cf. Anderson, 1984, p. 159):

$$\text{deviance difference} = N \ln\left(1 + \frac{1}{N-1}T^2\right). \tag{8}$$

(This is proved by Anderson for the log likelihood ratio defined for the contrast vectors. For complete data, the transformation to the contrasts entails no difference for the likelihood ratio.)

5.2. WALD TESTS

The Wald test can be calculated for θ consisting of $r = p - 1$ linearly independent contrast vectors, e.g., difference contrasts $\theta_j = \mu_{j+1} - \mu_j$ for $j = 1, \ldots, p-1$. The contrasts now refer to the parameters instead of the observations, which is the reason why calculation of the observation contrasts D_{ij} is unnecessary. For parameters estimated by ML, the Wald statistic produced for complete data is equal to $W = \{N/(N-1)\}T^2$. The Wald statistic based on REML estimates here is equal to T^2 . The REML Wald test may be expected to have a closer approximation to the type I error probability than the ML Wald test.

For complete data, this gives us three tests: the asymptotic chi-squared likelihood ratio test, the asymptotic chi-squared Wald test, and the exact Hotelling's test. The three tests are closely related, which is seen as follows. A first-order Taylor series of the function (8) shows that the deviance difference can be approximated by

$$\text{deviance difference} \approx \frac{N}{N-1}T^2 = \frac{N}{N-p+1}(p-1)F.$$

(This approximation is poor if $T^2/(N-1)$ is large.) F denotes the exact F-transform of the T^2 statistic. When N tends to infinity, the factor $N/(N-p+1)$ tends to 1 and the distribution of $(p-1)F$ tends to a chi-squared distribution with $p-1$ degrees of freedom. This implies that the three tests will coincide for large sample sizes, unless $T^2/(N-1)$ is rather large (but then all tests will yield very small p-values anyway). The tests all are functions of the same test statistic T^2, but use different rejection regions. Since Hotelling's test is exact, it is to be preferred.

Example 2
The example about salsolinol excretion is continued without taking into account the two groups. The fully multivariate ("*MANOVA*") approach is used.

For the test of the effect of "day", Hotelling's T^2 yields $F(3, 11) = 1.55$ ($p = 0.26$) while the Wald test based on ML estimation is $W = 5.92$ (*d.f.* = 3, $p = 0.12$) and the Wald test based on REML estimation is $W_R = 5.50$ ($p = 0.14$). This is in accordance with the formulae above with $N = 14$, $p = 4$. Although the "day" effect is non-significant by either test, this example does illustrate that using the exact F-test rather than the asymptotic chi-squared test is an important modification for this relatively small sample size.

5.3. INCOMPLETE DATA

The use of the multilevel approach has an important advantage in the case of incomplete data. Recall that it is assumed throughout this paper that the missing data are MAR or MCAR. Parameter estimates and the Wald test statistic can still be calculated by software for random coefficient models when data are incomplete, without any change in model specification or setup. The Wald test still is valid asymptotically (when for each treatment condition, the number of available cases is large). However, referring the Wald statistic to the chi-squared distribution does not take into account the fact that the covariance matrix is estimated so this asymptotic approximation will be liberal in small samples.

Continuation example 2
For each of 8 subjects one measurement was deleted. The deleted measurements were distributed evenly over the 4 time points. The REML version of the Wald test yields $W_R = 4.96$. With *d.f.* = 3, this result has $p > 0.20$ in a chi-squared distribution. This illustrates that the analysis still is possible, although the test is valid only asymptotically (and therefore not very accurate for this small sample size), and the loss of data leads to a loss of power.

6. Multivariate Tests in between-within Designs

A between-within design (e.g., Maxwell and Delaney, 1990, Chap. 14; Stevens, 1996, Chap. 13) contains at least one between-subjects factor (grouping or classification variable for the subjects) and at least one within-subjects factor (a classification for the measurement occasions). We consider a design with one between-subjects and one within-subjects factor, and refer to these as "group" and "condition", respectively.

The overall multivariate null hypothesis is that the groups have identical vectors of population means. This is decomposed into two sub-hypotheses. The first is the hypothesis of no main group effect: the sum score over the repeated measures has the same population mean in the various groups. The second is the hypothesis

of no group by condition interaction: the differences between conditions have the same population means in the various groups. The group by condition interaction usually is the most important for the research question investigated. For incomplete data, the tests for the main group effect and the group by condition interaction effect need no longer be orthogonal, as they are in the case of complete data. This is analogous to the situation of two-way between-subjects designs with unequal numbers of subjects per cell, cf. Kleinbaum et al. (1988, Chap. 20).

Multivariate tests of these hypotheses, for complete as well as for incomplete data, can be carried out in the multilevel framework according to the same procedure as indicated above. We focus on Wald tests for the fully multivariate model (without a restriction on the covariance matrix), which are equivalent to Hotelling T^2 tests for the designs for which this test is available. The procedure can be summarized as follows:

1. fit a model where the fixed part expresses the alternative hypothesis, the random part at level 1 is empty, and the random part at level 2 incorporates correlated random slopes of p variables (e.g., polynomial or dummy codings) representing the p conditions;
2. express the null hypothesis as a set of linear constraints on the vector of fixed parameters, and test these by a Wald statistic.

6.1. MULTILEVEL FORMULATION OF BETWEEN-WITHIN DESIGNS

In the usual repeated measures notation the dependent variable is denoted by Y_{ijk}, where k denotes the group. We find it more convenient here to employ the multilevel usage where the number of indices of the dependent variable reflects the number of hierarchical levels. The grouping variable here is a categorical variable defined for the subjects, not a level of nesting in the multilevel sense (the latter representation would imply a random group effect, whereas the hypothesis now is about fixed group effects). Subject number i runs from 1 to N, and by $g(i)$ we denote the group of subject i. The number of groups is m. We define the dummy variable w_k by $w_{ki} = 1$ if subject i is in group k (i.e., $g(i) = k$) and $w_{ki} = 0$ otherwise.

The model for m groups with p conditions now can be expressed as

$$Y_{ij} = \mu_{g(i),j} + U_{ij} = \sum_{k=1}^{m} \sum_{h=1}^{p} \mu_{kh} w_{ki} z_{hj} + \sum_{h=1}^{p} U_{ih} z_{hj}, \tag{9}$$

where μ_{kh} denotes the population mean under condition h for group k, while the occasion dummies z_{hj} are defined by (6) and the meaning of the random effects U_{ih} is as above. The random part is the same as that of model (4), and represents an unrestricted homoscedastic covariance matrix. The fixed part is formed by the mp product dummy variables $w_k z_h$.

The three hypotheses can be represented as linear constraints on the parameters μ_{kh} of the fixed part. The overall hypothesis of equality of the m populations is given by

$$\mu_{k+1,h} - \mu_{kh} = 0 \quad (k = 1, \ldots, m-1; h = 1, \ldots, p). \tag{10}$$

The hypothesis that there is no main effect of group can be formulated as

$$\sum_{h=1}^{p} (\mu_{k+1,h} - \mu_{kh}) = 0 \quad (k = 1, \ldots, m-1). \tag{11}$$

The hypothesis of no group by condition interaction is

$$\mu_{k+1,h+1} - \mu_{k+1,h} - \mu_{k,h+1} + \mu_{kh} = 0$$

$$(k = 1, \ldots, m-1; h = 1, \ldots, p-1). \tag{12}$$

The restrictions (11) and (12) jointly are equivalent to (10).

6.2. TESTS

In the case of two groups and complete data, hypothesis (10) can be tested by Hotelling's two-sample T^2 test (e.g., Stevens, 1996, Chap. 4), while hypothesis (12) can be tested by the same test applied to the vectors of contrasts D_i defined above. Hypothesis (11) can be tested for complete data, for an arbitrary number of groups, by a univariate F test applied to the sum scores. For complete data and more than two groups, hypotheses (10) and (12) are tested by multivariate F tests, in some cases exact and otherwise carried out using tables or very good approximations (e.g., Anderson, 1984, Chap. 8; Maxwell and Delaney, 1990, Chap. 14; Stevens, 1996, Chap. 5; Tatsuoka, 1988, Chap. 8).

For complete data, Wald tests are equivalent to certain of these F tests. This is of practical importance because Wald tests can be calculated routinely by multilevel software. To elaborate this, we use the transformations of T^2 statistics to F distributions given in the mentioned literature. The Wald test statistic for the multivariate null hypothesis (10) is denoted by $W(\text{multi})$, for null hypothesis (11) of no group main effect by $W(G)$ and for null hypothesis (12) of no group by condition interaction by $W(G \times C)$. The corresponding Wald test statistics based on REML estimates are denoted using W_R instead of W.

The F statistic for the main effect of group – null hypothesis (11) – can, in the case of an arbitrary number m of groups and complete data, be expressed as

$$F(G) = \frac{N-m}{N(m-1)} W(G) = \frac{1}{m-1} W_R(G). \tag{13}$$

For the multivariate tests, first consider $m = 2$ groups. For the multivariate test of equality of the two vectors of population means, the T^2 test statistic for the complete data case is again a multiple of the Wald statistic,

$$T^2(\text{multi}) = W_R(\text{multi}) = \frac{N-2}{N} W(\text{multi}) \tag{14}$$

and can be transformed to an exact F-distribution by

$$F(\text{multi}) = \frac{N-p-1}{(N-2)p} W_R(\text{multi}). \tag{15}$$

This statistic has, under the null hypothesis, the F-distribution with p and $N-p-1$ degrees of freedom. This test can also be used for p-dimensional multivariate data without a repeated measures structure.

The situation for the group by condition interaction closely parallels the test of the hypothesis of multivariate equality of means. For $m = 2$ groups with complete data, the Wald and T^2 statistics are related by

$$T^2(G \times C) = W_R(G \times C) = \frac{N-2}{N} W(G \times C) \tag{16}$$

and transformed to an exact $F(p-1, N-p)$ distribution by

$$F(G \times C) = \frac{N-p}{(N-2)(p-1)} W_R(G \times C). \tag{17}$$

Continuation example 3
We continue the example of the salsolinol excretion data. There are $m = 2$ groups and $p = 4$ measurements, so in representation (9) there are 2 dummy variables w_{ki} and 4 dummy variables z_{hij}, leading to 8 product variables $w_{ki}z_{hij}$.

For the test of the main group effect, $F(1, 12) = 2.48$ ($p = 0.14$). It follows from (13) and $m = 2$ that this F statistic is equal to the Wald statistic $W_R(G)$. For the multivariate test of equality of the 2 group means, $F(4, 9) = 1.101$ ($p = 0.41$) while W_R (multi) = 5.87. This corresponds with (15). For the group by condition interaction, $F(3, 10) = 0.197$ ($p = 0.90$) while $W_R(G \times C) = 0.71$, in accordance with (17). None of the tests leads to a significant result.

The situation is more complicated for testing multivariate equality of the group means, or group by condition interaction, for more than 2 groups. These testing problems can be treated simultaneously by defining $r = p$ for the former and $r = p - 1$ for the latter hypothesis. There are various multivariate F-tests for the complete data case (see, e.g., Anderson, 1984 or Tatsuoka, 1988). It was proven by Kleinbaum (1973) that, of these tests, it is the Lawley–Hotelling trace L (also called the Hotelling trace) that corresponds to the Wald test. More specifically,

$$L = \frac{1}{N} W = \frac{1}{N-m} W_R.$$

In this context, $(N - m)L = W_R$ is also called the Hotelling T_0^2 statistic. Generally valid approximations of its null distribution by a transformed F-distribution are not available, but special approximations and tables exist for its null distribution (cf. Anderson, 1984, and Pillai, 1983). The asymptotic distribution of W_R is chi-squared with $r(m - 1)$ degrees of freedom.

7. Summary and Discussion

In the statistical treatment of repeated measures, an important distinction is between the traditional mixed model approach, also called the univariate approach, and the multivariate (*MANOVA*) approach (e.g., Maxwell and Delaney, 1990; Stevens, 1996). The multilevel (or hierarchical linear model) approach, mentioned by Goldstein (1995), steers a middle road between the traditional mixed model, which includes only the main subject effect as a random effect, and the multivariate approach, which makes no restrictions on the covariance matrix except the requirement of homoscedasticity in the multi-group case. The multilevel approach is based on random coefficients of linear and non-linear functions of the within-subject variables, and implies assumptions for the covariance matrix of the repeated measures which are intermediate between the very strict assumptions of the traditional mixed model and the very loose assumptions of the multivariate approach.

Since the multilevel approach allows the researcher to select a parsimonious and well-fitting model for the covariance matrix, the associated tests of fixed effects may be expected to have good power properties (cf. Reinsel, 1984 and Ware, 1985). Further investigations are necessary to study the importance of this gain in power, and to compare these tests with the epsilon-adjusted tests defined in the mixed model framework (as explained, e.g., by Maxwell and Delaney, 1990).

The usual implementation of the traditional approaches does not allow incomplete data. The multilevel approach, which includes the traditional mixed model and the fully multivariate model as boundary cases, allows incomplete data without any problems and can be implemented by multilevel software like *HLM* and *MlwiN* and by mixed model software like *SAS Proc Mixed*. Fixed effects are tested in multilevel analysis usually by means of deviance (i.e., likelihood ratio) or Wald tests. These tests are valid for large sample sizes. The advantage of the *MANOVA* approach is that under the normality assumption it yields exact tests also for small sample sizes. We indicated how deviance tests and Wald tests provided by multilevel software can be modified so that they are equal to these exact tests in the case of complete data.

Finally, we would like to signal an extension of this model. The multilevel formulation easily allows heteroscedastic models for multi-group data. This is done by letting the covariance matrices of the random slopes U_{ih} depend on the groups (cf. the remarks in Goldstein (1995) about complex variation and Snijders and Bosker (1999, Chap. 8), about heteroscedasticity). Thus a statistical treatment of heteroscedastic models for within-between designs is obtained which can be

applied to complete as well as incomplete data. Due to the asymptotic nature of the deviance and Wald tests, this procedure currently is useful mainly for intermediate and large sample sizes.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley and Sons.

Berk, K. (1987). Computing for incomplete repeated measures. *Biometrics* 43: 385–398.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park: Sage.

Dixon, W. J. (1992). *BMDP Statistical Software Manual*, Vol. 2. Berkeley, Los Angeles: University of California Press.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd edn. London: Edward Arnold.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yanh, M., Woodhouse, G. & Healy, M. (1998). *A User's Guide to MlwiN*. London: Multilevel Models Project, Institute of Education, University of London.

Hand, D. J. & Taylor, C. C. (1987). *Multivariate Analysis of Variance and Repeated Measures*. London: Chapman and Hall.

Hays, W. L. (1988). *Statistics for the Social Sciences*, 4th edn. London etc.: Holt, Rinehart and Winston.

Jennrich, R. I. & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 42: 805–820.

Kleinbaum (1973). Testing linear hypotheses in generalized multivariate linear models, *Communications in Statistics* 1: 433–457.

Kleinbaum, D. G., Kupper, L. L. & Muller, K.E. (1988). *Applied Regression Analysis and Other Multivariable Methods*, 2nd edn. Boston: PWS-KENT Publishing Company.

Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics* 38: 963–974.

Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. (1996). *SAS System for Mixed Models*. Cary, NC: SAS Institute Inc.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Maxwell, S. E. & Delaney, H. D. (1990). *Designing Experiments and Analyzing Data*. Belmont, CA: Wadsworth Publishing Company.

O'Brien, R. & Kaiser, M. K. (1985). *MANOVA* method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin* 97: 316–333.

Pillai, K. C. S. (1983). Hotelling's trace. In: S. Kotz, S. L. Johnson and C. B. Read (eds.), *Encyclopedia of Statistical Sciences*, Vol. 3. New York: Wiley, pp. 673–677.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd edn. New York: Wiley.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. & Congdon, R. T. (2000). *HLM5: Hierarchical Linear and Nonlinear Modeling*. Chicago: Scientific Software International.

Reinsel, G. C. (1984). Effects of the estimation of covariance matrix parameters in the generalized multivariate linear model. *Communications in Statistics – Theory and Methods* 3: 639–650.

Rogosa, D., Brandt, D. & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin* 92: 726–748.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

Schwertman, N. C. (1978). A note on the Greenhouse–Geisser correction for incomplete data split-plot analysis. *Journal of the American Statistical Association* 73: 393–396.

Snijders, T. A. B. & Bosker, R. J. (1999). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Snijders, T. A. B. & Maas, C. J. M. (1996). Using MLn for repeated measures with missing data. *Multilevel Modelling Newsletter* 8(2): 7–10.

Stevens, J. (1996). *Applied Multivariate Statistics for the Social Sciences*, 3d edn. Hillsdale: Lawrence Erlbaum Associates.

Tatsuoka, M. M. (1988). *Multivariate Analysis*, 2nd edn. New York: Wiley.

Van Der Leeden, R. (1998). Multilevel analysis of repeated measures data. *Quality and Quantity* 32: 15–29.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statisticia* 39: 95–101.